图书情报工作
第 65 卷 第 12 期 2021 年 6 月

融合多维特征的学术文献下载行为预测研究*

■ 谢豪 吴雪华 陈茜 唐晶 白云 毛进
武汉大学信息资源研究中心 武汉 430072

摘 要：[目的/意义]学术文献下载行为是科研人员文献检索行为的重要一环,对其预测的研究有助于深度理解科研人员检索行为,为学术资源检索平台优化检索结果、重构排序提供依据,从而提升检索系统的服务质量。[方法/过程]构建用户学术文献下载行为的多维特征体系,在机器学习算法基础上构造基于查询相关性和基于用户行为的子分类器,并采取加权策略构建学术文献下载行为预测混合模型。[结果/结论]实验结果表明,随机森林算法在两种分类器上均取得最佳性能;相较于仅基于查询相关性特征训练的模型,混合模型的准确率提高了 2.3%,F1 值提升了 1.3%。在混合模型中,基于用户行为的子分类器拥有更高权重;"下载量""是否采用专业/高级检索"和"发表时间"特征的贡献度较大。

关键词：文献下载预测 多维特征 机器学习 混合模型
分类号：G203
DOI：10.13266/j.issn.0252－3116.2021.12.011

学术资源检索平台作为科研人员获取学术信息的重要渠道,具有资源丰富、更新及时、获取便捷等优势。然而,学术资源的迅速增长也带来了信息过载的问题,导致检索成本增加,占用科研人员大量的时间和精力。针对检索结果重构排序以优化检索功能,对于提升学术资源检索平台服务的满意度、满足科研用户学术信息需求至关重要。

学术文献下载是科研用户学术检索的后续流程,理解学术文献下载行为能够为学术检索结果排序提供依据。当前学术文献下载研究主要集中在文献被引量和下载量的相关性上[1－3],倾向于将下载量作为文献计量评价指标,用以弥补文献被引量的时滞性问题。部分学者从知识产权的角度对过量下载行为的特点[4－5]、检测方法[6]进行探析,并提出相应对策[7]。但目前关于学术文献下载行为预测的研究较少,且仅停留在学术文献下载量预测层面[8],忽视了科研用户在学术检索时的信息交互行为所反映的用户偏好[9],未能从更细粒度的角度将科研用户的单次检索信息结合到下载预测中。而检索信息可以很大程度上反映用户的信息需求、检索目标与检索动机[10],因此探究单次检索中学术文献下载行为的影响因素,对于明晰用户检索意图、优化学术文献检索结果排序、提高科研人员检索效率和学术资源的利用率具有重要意义。

基于此,笔者提出一种融合多维特征的学术文献下载行为预测模型。在构建用户学术文献下载行为相关特征体系的基础上,采用机器学习算法建立基于查询相关性和基于用户行为的子分类模型,并采取加权策略构建混合分类模型用于预测用户的学术文献下载行为。

1 相关研究

学术检索能够实现对学术信息的过滤和筛选,满足学者的多元化需求和个性化兴趣。目前,学者主要从查询意图、查询式特征、检索策略等方面研究学术检索行为。查询意图指用户在检索过程中可能的潜在目的,可分为信息类、导航类和事务类[11]。基于不同查询意图进行的检索能够体现用户的个性化差异[12]。M. Khabsa 等[13]根据学术检索行为的特点,将学术用

* 本文系国家自然科学基金创新研究群体项目"信息资源管理"(项目编号:71921002)和国家自然科学基金青年项目"基于学术异质网络表示学习的知识群落发现"(项目编号:71804135)研究成果之一。

作者简介：谢豪(ORCID:0000－0002－1788－0468),硕士研究生;吴雪华(ORCID:0000－0003－0231－2975),硕士研究生;陈茜(ORCID:0000－0003－1640－7270),硕士研究生;唐晶(ORCID:0000－0002－1211－5812),硕士研究生;白云(ORCID:0000－0002－7590－1263),硕士研究生;毛进(ORCID:0000－0001－9572－6709),副教授,博士,通讯作者,E-mail:danveno@163.com。

收稿日期:2020－11－26 修回日期:2021－03－03 本文起止页码:112－121 本文责任编辑:徐健

户的查询意图分类为导航类和信息类,其中导航类查询意图指用户的目标为特定学术文献,信息类查询意图指用户希望获取某一主题的相关信息[14]。作为查询式的基本特征,查询式的构造能揭示用户最直接的需求[15]。X. Li 等[16]通过分析查询式内容发现学术检索中多为实体检索,这些实体能够反映用户感兴趣的主题。根据人类信息行为理论[17],学术检索行为可以归纳为研究探索性、任务导向性和技巧依赖性 3 种类型。不同类型的学术检索行为对应的检索策略有所不同[18],如果用户希望了解领域内的发展态势,则需要获取大量文献,检索行为属于研究探索性,关键词检索、期刊检索等为首选的检索策略;而当用户有较为明确的学术检索目标时,检索行为具有任务导向性,倾向于使用精确匹配模式。

学术文献下载是学术检索的后续环节,学者开展研究时通过检索文献寻找领域中与当前工作相关的学术信息,并从检索结果中下载符合期望的文献。对学术文献下载的研究一方面可以弥补被引频次无法反映隐形引用文献的学术价值的缺点[19],即考虑被阅读过但未被引用的文献的贡献;另一方面,由于论文从完成到被学者引用,需要经历出版机构的评审、读者的理解等步骤,导致引文分析存在一定的延时[20]。而对文献下载的研究能够缓解这种滞后性[21],可以较快反映论文的价值[22]。相比于引文数据,下载量更具测评区分度与敏感性,在统计学上的数值特征与引文有所不同[23]。因此对于文献下载的研究,可以作为引文行为的补充,为文献学术影响力的研究提供新的视角。

作为文献被使用的一个指标,历史下载量能及时反映论文被使用的情况,在一定程度上早于被引发现论文的引用价值[24]。虽然单篇开放获取论文的下载频次与被引频次之间的相关性并不明显[25],但是《国际会计信息系统》期刊中论文被引用次数与论文进入下载量前 25 位的次数却显著相关[26],说明文献下载频次与引用频次之间的关系在单篇论文层次和期刊层次的相关性存在差异。因此,利用论文发表一段时间之后的下载量可以对论文和期刊未来的引用量进行较为准确的预测[22]。此外,过量下载成为高校图书馆普遍面临的问题,徐文贤等[27]通过调查国内外的过量下载案例,发现科研学术需求和商业利益是造成过量下载的主要原因。

综上所述,现有研究大多通过分析下载量与引用量的关系以评估文献的学术影响力,或者分析过量下载现象以规范对学术资源检索平台的使用,缺乏对学术检索中用户下载行为决策的探究。因此,笔者从查询相关性和用户行为两种视角出发,构建学术检索中的用户文献下载行为预测模型。

2 融合多维特征的学术文献下载行为预测模型

2.1 问题定义

笔者将单次学术检索中的文献下载行为预测定义为一个二分类问题:给定用户 u 及文献检索结果 $D=\{d_1,d_2,d_3,\cdots\cdots,d_n\}$（n 为检索结果文献总数),对于 D 中的任一文献 d_i,预测用户 u 是否会进行下载。预测标签 $y_i\in\{0,1\}$,其中 1 代表下载,0 代表未下载。

2.2 学术文献下载行为预测模型框架

学术文献下载行为本质上受到用户需求的驱动,其背后的基本假设是与用户需求越匹配的学术文献,越可能被下载。基于以往研究,笔者认为这种需求的匹配程度体现在两个方面:一是语义层面的检索式与学术文献的相似性,二是用户行为反映出的用户需求与文献之间的相似关系。据此,首先针对这两类信息,分别构建基于查询相关性的子分类器和基于用户行为的子分类器,然后进一步整合这两个子分类器提出混合分类器。基于查询相关性的子分类器主要对文献特征和用户查询式特征进行学习,其目的是通过文献与用户检索需求的匹配程度来预测下载行为;基于用户行为的子分类器借助 item2vec 模型[28]从用户行为记录中提取文献嵌入表示,旨在挖掘文献之间的潜在关联;混合分类器对上述两个子分类器的预测结果进行加权,以全面捕捉学术文献下载行为的影响因素,提升模型效果。

2.2.1 基于查询相关性的子分类模型

查询相关性反映检索结果和用户需求的匹配程度,直接影响用户的浏览、下载、利用等后续行为,是一种重要的检索结果排序依据[29-30]。查询相关性的评估涉及文献和用户查询意图两类特征。

文献特征主要从质量和内容两个层面衡量文献能否满足用户需求。一方面,信息质量显著影响用户对信息的有用性认知,进而影响其态度和行为决策[31]。反映在学术信息搜寻场景中,高质量的学术文献能提升用户的感知有用性,进而促使下载行为的产生。常用的文献质量衡量指标包括被引量、下载量、来源期刊和发表时间[2,32]。其中,被引量和下载量能用于衡量文献的影响力,而来源期刊、发表时间则分别反映了文

献的可靠性和时效性。另一方面,文献内容满足用户信息需求的程度也影响着用户的下载行为决策,可通过计算文献内容与当前查询式的匹配度来衡量[33]。

查询意图反映了用户的查询目标和动机,影响着用户后续的浏览、下载等行为选择[34]。例如,在学术信息搜寻情境下,用户的目标既有可能是获取特定文献(即导航类查询意图),也有可能是了解某一主题、机构或作者的发文数量等信息(即信息类查询意图),前者比后者更容易产生文献下载行为。鉴于查询意图较为抽象,难以直接识别,部分文献通过查询式构造特征来间接反映[13,34-35]。参考已有文献,笔者选取查询式长度、是否为题名、检索字段、是否采用精确匹配、是否采用专业/高级检索 5 个查询式特征。

表 1 总结了基于查询相关性的子分类模型中的特征体系,具体如下:

表 1 基于查询相关性的子分类模型的特征体系

类型	具体特征	描述
文献特征	来源期刊	是否来自权威期刊
	发表时间	-
	被引量	-
	下载量	-
	文献匹配度	文献与用户信息需求的匹配程度
用户查询式特征	查询式长度	查询式中的词语数量
	是否为题名	查询式是否为文献的完整题名
	检索字段	查询式是否包含题名、DOI 或作者字段
	是否采用精确匹配	是否要求检索词与文献某一字段完全匹配
	是否采用专业/高级检索	是否包含高级检索或专业检索运算符

各特征的提取方法如下:

(1)文献特征提取。文献元数据特征可从文献著录信息中直接获取。其中,来源期刊特征主要判断文献是否来自权威期刊,即被北大核心、中文社会科学引文索引(CSSCI)、中国科技论文统计源期刊(CST-PCD)、工程索引(EI)、科学引文索引(SCI)等收录的期刊。发表时间特征为文献发表年份减去当前浏览或下载年份。由于文献被引量和下载量的分布具有极差大、不均匀的特点,因此笔者采用等频分箱进行处理,使每个区间内包含大致相等的样本数量。对于被引量,分箱后包含 3 个区间:低被引量区间为[0,1]、中被引量区间为(1,5]、高被引量区间为(5,+∞)。下载量数据分箱后也包含 3 个区间:低下载量区间[0,27]、中下载量区间(27,95]、高下载量区间(95,+∞)。

在计算文献匹配度特征时,考虑到一个查询式中可能存在多个检索字段,因此分成如下两部分进行:①对于内容相关的检索字段(如主题、题名等),采用关键词匹配法计算检索词与文献内容的匹配程度。具体做法为:对查询式、文献标题和摘要进行分词,之后计算出现在文献标题、摘要和关键词中的检索词个数占检索词总数的比例,以此作为匹配度。例如:查询式包含 3 个词汇,其中 2 个出现在文献的标题、关键词或摘要中,则匹配度为 0.66。②对于元数据相关的检索字段(如作者、期刊等),若查询式中存在与文献完全匹配的检索字段,则在内容匹配度上添加一个 0 到 1 之间的附加值,具体取值作为超参数,在模型训练时进行调优。公式表示如下:

$$score_{total} = \begin{cases} score_{content} & \text{不存在相匹配的元数据字段} \\ score_{content} + \beta \ (\beta \in [0,1]) & \text{存在相匹配的元数据字段} \end{cases} \quad \text{公式(1)}$$

其中,$score_{total}$ 为总体匹配度,$score_{content}$ 为内容匹配度,β 为附加值。

(2)用户查询式特征提取。查询式长度通过分词并计算词语个数来获取,提交较长查询的用户通常更有可能搜索具体的信息;是否为题名的判断标准如下:若查询式和文章题名的匹配词数大于 5,认为查询式为题名,否则判定为非题名。如果用户直接检索题名,表明用户更有可能明确查找特定的学术文献;检索字段特征主要判断查询式中是否包含作者、DOI 或题名字段,包含上述字段的查询式更有可能代表明确的查询意图;精确匹配、高级检索和专业检索通过正则表达式进行识别,其中包含精确匹配的查询式中通常带有精确匹配符(引号等);高级检索和专业检索通常包含逻辑运算符“and”“or”“not”“*”“+”“^”或限定检索顺序运算符“(”。通常不同的检索策略能在一定程度上反映用户的意图、偏好及信息需求类型。

2.2.2 基于用户行为的子分类模型

推荐领域相关研究指出,用户浏览或购买商品的行为序列中蕴含着商品的相似性信息。基于用户行为序列训练的商品嵌入表示能够将原始的高维稀疏数据映射到低维的特征空间,使相似的商品在空间距离的度量上相近,从而建模商品间的潜在联系,提升推荐效果[36-38]。笔者借鉴上述思路,基于用户与文献的交互行为记录,训练文献的低维嵌入表示作为子分类模型的特征,旨在捕获文献之间的深层关联。

推荐领域训练商品嵌入表示的重要思路之一是借

鉴自然语言处理中的词向量表示模型 word2vec[28]。具体做法是将商品视为 word2vec 模型中的单词，将用户浏览或购买商品的集合视为 word2vec 模型中的单词序列(句子)，将出现在同一个集合的商品对视为正样本，利用带负采样的 Skip-gram 模型(Skip-gram with Negative Sampling, SGNS)学习商品的低维嵌入表示(item2vec)。笔者将文献视为单词，将每个用户在一定时间段内浏览的文献集合视为一个句子。具体而言，用 article_id 字段标识一篇文献，则某个用户的所有文献浏览记录可表示为：

$$user_i = [article_id_1, article_id_2, \cdots, article_id_m] \quad \text{公式(2)}$$

其中，i 表示第 i 个用户，m 表示该用户浏览的文献数量。

用于训练 item2vec 模型的数据表示为：

$$train_list = [user_1, user_2, \cdots, user_i, \cdots, user_n] \quad \text{公式(3)}$$

其中，n 代表用户数。

将上述的用户浏览文献集合输入 SGNS 模型，学习文献的低维嵌入表示：

$$item_embedding = SGNS(train_list) \quad \text{公式(4)}$$

之后将得到的文献嵌入表示输入分类器，输出子分类模型的预测结果。

2.2.3 混合分类模型

混合分类模型对两个子分类模型预测的下载/未下载概率进行加权，得到最终的预测结果。权重系数在模型调优时确定。

综上，笔者提出的学术文献下载行为预测模型整体框架如图 1 所示，由以上 3 个分类器组合构成。对于输入的用户检索和浏览行为记录，该模型处理流程如下：首先，采用机器学习算法分别构建基于查询相关性和基于用户行为的子分类器；其次，基于两个子分类器构建混合分类器；最后，根据混合分类器的输出结果预测用户是否会下载某篇文献。

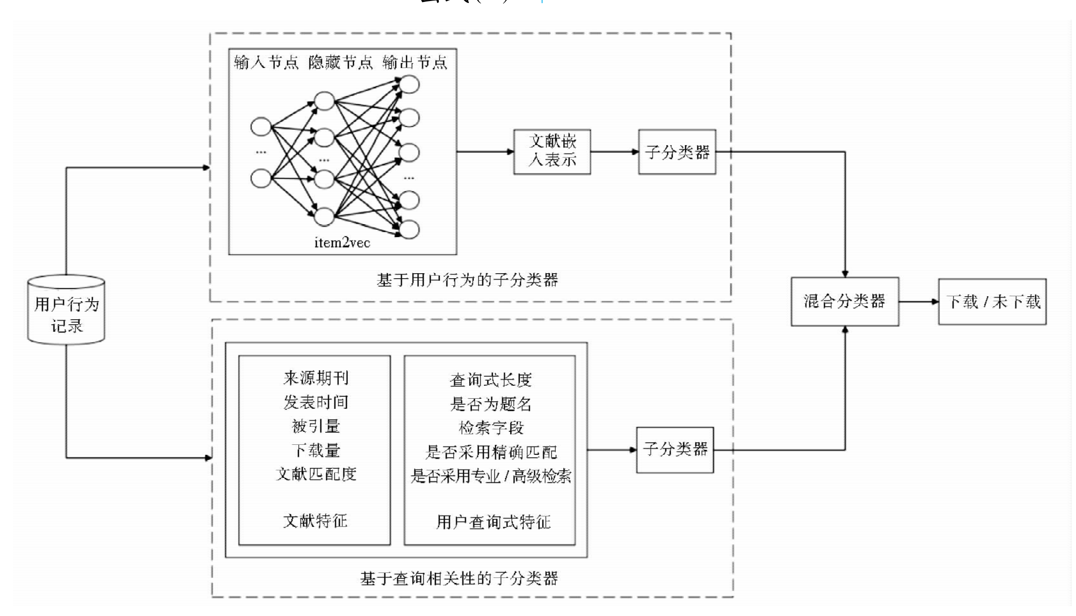


图 1 学术文献下载行为预测混合模型框架

3 实验与结果分析

3.1 实验设计

实验过程整体框架见图 2。

分为以下环节：

(1)数据字段扩充：为满足研究需求，编写爬虫采集文献和期刊相关字段的信息，对源数据进行扩充。

(2)数据预处理：对扩充后的数据集进行数据关联、删除无效数据等预处理，得到可用于实验的数据，之后按照 8:2 的比例划分为训练集和测试集。

(3)特征抽取：一方面，按照 2.2.1 中的方法对训练集中所有数据进行文献特征和用户查询式特征的抽取；另一方面，按照 2.2.2 所示方法，使用数据集中非机构用户的浏览行为数据训练 item2vec 模型，提取基于用户浏览行为记录的文献嵌入表示。

(4)模型训练：使用环节 3 中的训练数据，分别训练基于查询相关性和基于用户行为的子分类器。目前，在分类问题中较为成熟的机器学习算法有逻辑斯蒂回归(Logistic Regression, LR)、决策树(Decision Tree, DT)、朴素贝叶斯(Naive Bayes, NB)、支持向量机

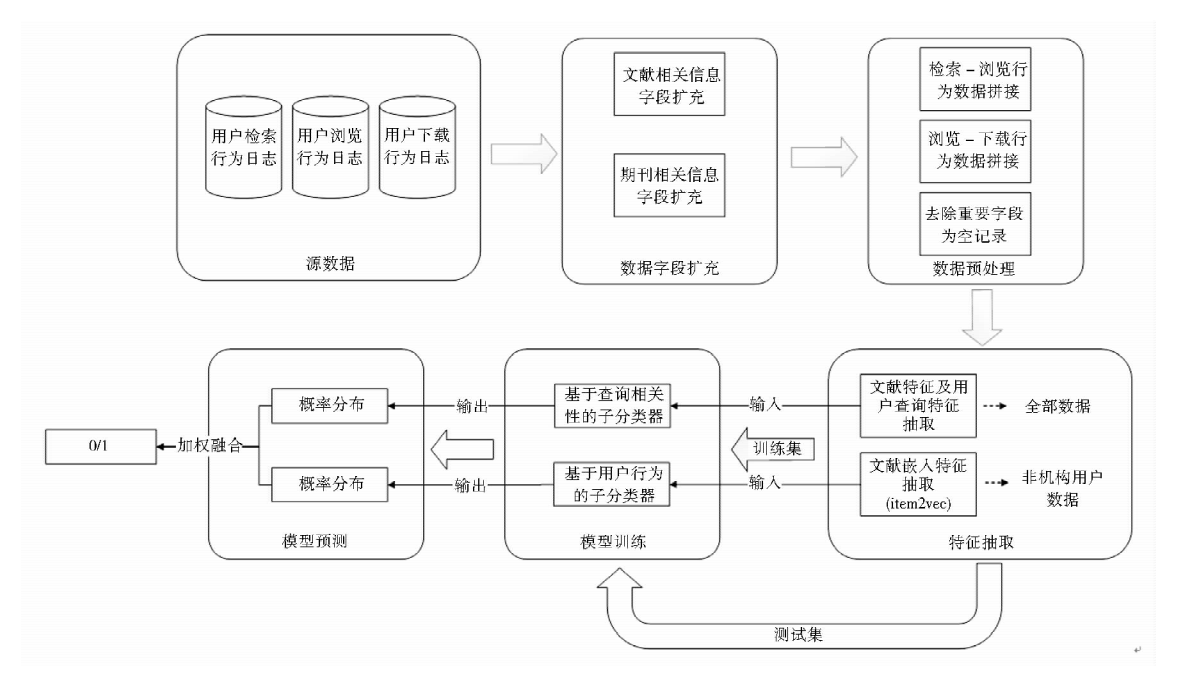


图 2　实验流程

(Support Vector Machine, SVM)、多层感知机(Multi-Layer Perceptron, MLP)以及随机森林(Random Forest, RF)。笔者使用上述分类算法进行实验,选用性能最佳的算法构建最终的混合分类器。

(5)模型预测与融合:使用同样的方式提取测试集数据的特征,将测试集中抽取的文献特征和用户查询式特征输入到基于查询相关性的子分类器中得到预测结果 p_i^p,将测试集中提取的文献嵌入表示输入基于用户行为的子分类器中得到预测结果 p_i^u,最后将两个预测结果进行加权融合,构建混合分类器,表示为:

$$p_i = \begin{cases} \alpha * p_i^p + (1-\alpha) * p_i^u, & i\ \text{为非机构用户行为产生的文献数据记录} \\ p_i^p, & i\ \text{为机构用户行为产生的文献数据记录} \end{cases} \qquad \text{公式(5)}$$

其中,α 为模型融合权重系数。

3.2 数据集与预处理

本文研究数据来自“慧源共享”全国高校开放数据创新研究大赛组委会提供的万方数据知识服务平台期刊文献用户行为日志数据集,包括用户检索行为日志 37 544 670 条、用户浏览行为日志 11 998 421 条以及用户下载行为日志 14 025 159 条,时间跨度为 2019 年 12 月 1 日至 2020 年 1 月 31 日[39]。由于源数据包含的字段信息无法满足研究需求,笔者对用户浏览行为日志数据集字段进行了扩充,具体做法是使用用户浏览行为日志中每条记录的文献题名字段和文献作者字段在万方数据知识服务平台进行高级检索,抓取检索结果排名首位的论文相关信息。扩充字段包括两个方面:文献元数据和期刊元数据。文献元数据字段包括摘要、发表年份、被下载数以及被引数;期刊元数据字段包括来源期刊 ID、期刊级别、期刊名称、期刊总下载量、期刊总被引量以及期刊影响因子。

用户在进行学术文献下载时通常有两种行为模式,如图 3 所示。一种是在浏览文献摘要等详细信息之后决定是否下载,即检索－浏览－下载/未下载;另一种则是直接通过检索列表中的粗略信息判断文献是否满足自己的需求,即检索－下载/未下载。用于本文实验的数据为用户检索后进行浏览的数据记录,正样本为用户检索－浏览－下载数据,负样本为用户检索－浏览－未下载数据。

用户行为日志源数据集存在的问题是,用户的检索、浏览、下载行为记录分表保存,且对于机构用户数据无法定位到个人,若只对非机构用户数据进行分析,则会造成巨大的数据资源浪费。为了将机构用户的行为记录也纳入到实验数据中,笔者采用如下方法对用户检索、浏览、下载行为日志进行拼接:①基于同一用户(user_id 字段值相同)对同一文献(article_id 字段值相同)的浏览时间与检索时间的最小时间差将检索行为与浏览行为记录进行拼接;②对检索词以及标题关

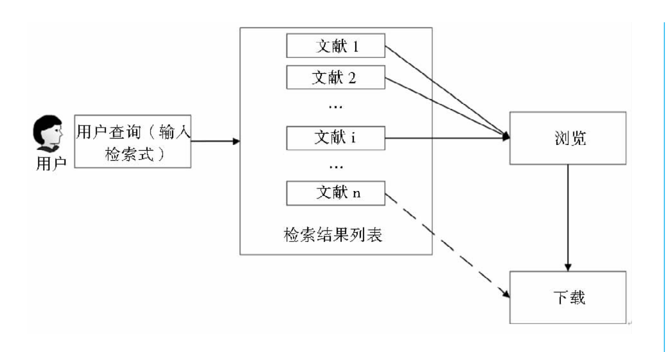


图 3　用户学术下载行为模式

键词进行分词、去除停用词,使用关键词共现法判断检索词与检索文献的相关性。具体做法是判断检索词的分词结果列表与标题关键词的分词结果是否存在共现词,若存在则认为相关,否则认为不相关,删除不相关的数据;③同样利用同一用户对同一文献的下载时间与浏览时间的最小时间差将浏览行为日志与下载行为日志中的记录进行拼接。去除文献发表年份及摘要为空值的数据后,得到最终用于实验的数据 2 383 933 条,具体如表 2 所示:

表 2　预处理后的数据(单位/条)

数据类别	检索－浏览－下载数据	检索－浏览－未下载数据	总数
机构用户数据	438 643	1 486 176	1 924 819
非机构用户数据	84 662	374 452	459 114
总数	523 305	1 860 628	2 383 933

从表 2 可以看到,检索－浏览－下载数据与检索－浏览－未下载数据的总数存在不平衡现象,比例大约为 1:3.5,因此在训练基于查询相关性的子分类器时,为了保证数据集中两类标签数据的平衡,笔者对检索－浏览－未下载数据样本进行下采样,从中随机抽取与检索－浏览下载数据等量的数据用于实验。

3.3　评估指标

笔者选用准确率(accuracy)、召回率(recall)、精确度(precision)以及 F1 值作为模型的评估指标,计算公式如下:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \qquad 公式(6)$$

$$Recall = \frac{TP}{TP + FN} \qquad 公式(7)$$

$$Precision = \frac{TP}{TP + FP} \qquad 公式(8)$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \qquad 公式(9)$$

其中,TP 是模型预测用户发生下载行为且用户确实发生的样本数,TN 表示预测用户不发生下载行为且事实上确实未下载的样本数,FP 表示预测用户发生下载行为但实际上未发生的样本数,FN 表示预测用户不发生下载行为但实际上发生的样本数。

3.4　实验结果分析

表 3、表 4 和表 5 分别展示了基于查询相关性的子分类器、基于用户行为的子分类器以及混合分类器的预测结果。

表 3　基于查询相关性的子分类器的实验结果

算法	accuracy	recall	Precision	F1
LR	0.629	0.651	0.622	0.636
DT	0.654	0.765	0.625	0.688
NB	0.599	0.739	0.577	0.648
SVM	0.630	0.665	0.622	0.638
MLP	0.659	0.781	0.626	0.695
RF	**0.666**	**0.807**	**0.628**	**0.706**

表 4　基于用户行为的子分类器的实验结果

算法	accuracy	recall	Precision	F1
LR	0.510	0.801	0.510	0.622
DT	0.648	0.672	0.646	0.659
NB	0.516	**0.993**	0.510	**0.674**
SVM	0.507	0.639	0.510	0.567
MLP	0.575	0.466	0.602	0.525
RF	**0.657**	0.443	**0.783**	0.566

表 5　混合分类器的实验结果

	accuracy	recall	Precision	F1
融合模型(基于 RF)	0.689	0.799	0.654	0.719

从表 3 和表 4 的实验结果可以看出,对于基于查询相关性和基于用户行为的子分类器,随机森林在各个评价指标上均取得最佳性能,因此笔者选用随机森林算法训练的分类器进行最终的模型融合。值得一提的是,在基于用户行为的子分类器中,随机森林的精确度超出性能表现第二的决策树模型 14 个百分点,达到 78.3%,但在召回率上过低;基于查询相关性的子分类器则刚好相反。将两种分类器进行融合能够起到互补作用,从而提升性能,表 5 的实验结果也证明了混合分类器的准确率和 F1 值优于子分类器,其中准确率相较基于查询相关性的子分类器提升了 2.3%,F1 值提升了 1.3%。

3.5　子分类器权重占比分析

为探究基于查询相关性的子分类器和基于用户行为的子分类器在混合分类器中的权重占比,笔者采用穷举法探索模型融合权重系数 α 的最优值,其中 $\alpha \in [0,1]$,步长为 0.1。实验结果如图 4 所示。从图 4 中

可以看出，α 值为 0.3 时，模型准确率达到最高，随着 α 值的继续升高，准确率开始降低；α 值为 0.4 时，模型 F1 值达到最高，随后随 α 值的升高而下降。综合考虑准确率和 F1 值，模型权重系数 α 的最优值为 0.4，此时准确率相对峰值差距较小，而 F1 值达到最优。这表明在性能最优的混合分类模型中，基于用户行为的子分类相较于基于查询相关性的子分类器占有更高的权重，这可能是因为基于用户行为的子分类器能充分考虑用户的历史行为记录，从更细粒度的角度抽取文献的内在特征、学习文献之间的潜在联系。

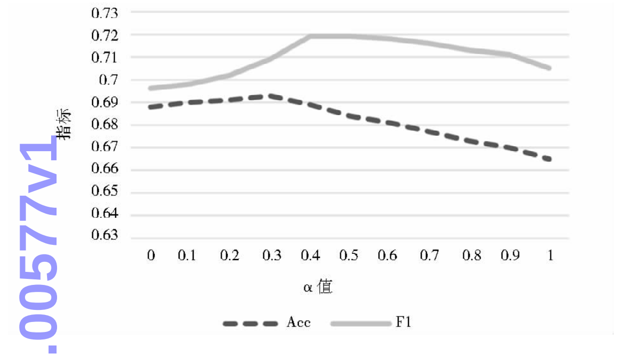


图 4　准确率、F1 值随权重系数 α 值的变化

3.6　特征贡献度分析

特征贡献度分析能够揭示特征对不同数据类别的区分能力，增强分类模型的可解释性，为之后的模型优化提供参考。针对查询相关性特征，笔者基于信息增益计算各特征在学术文献下载行为预测任务中的贡献度，结果如图 5 所示：

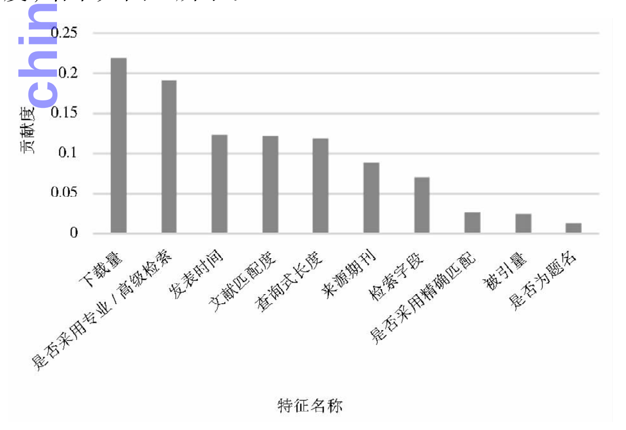


图 5　特征贡献度分布

根据特征的贡献度分布，笔者将贡献度大于 10% 的特征称为高贡献度特征，将贡献度低于 10% 的特征称为低贡献度特征。在高贡献度特征中，“下载量”作为文献质量的重要评价指标，对于样本能否正确分类具有明显影响。如图 6(a)所示，在下载量较小时，用户浏览后未下载的数据占比较大；在下载量较大时，用户浏览后下载的数据占比较大，说明科研人员下载文献时，倾向于根据文献下载量对文献质量做出判断，这符合信息资源分布中的马太效应。

“是否采用专业/高级检索”和“查询式长度”皆可反映用户的检索目标是否明确。当用户采用“专业/高级”检索，或者提交长查询式时，说明用户的检索目的非常明确，对于检索的结果要求较高，因此下载的可能性更低。而当用户采用较短的查询式时，他们的兴趣在于了解领域内的发展态势，需要获取大量文献，从而更可能产生下载行为。根据实验的结果来看（见图 6(b)和 6(e)），“是否采用专业/高级检索”和“查询式长度”可作为分类的重要特征。

“发表时间”反映了文献内容的新颖程度。实验数据表明，当发表时间距离当前查询日期较近时，用户浏览后下载的比例更高；当发表时间距离当前查询日期较远时，用户浏览后未下载的比例更高（图 6(c)）。说明科研人员在下载文献时，倾向于选择近期发表的文章。因此，“发表时间”对于预测结果具有一定的影响。

“文献匹配度”反映了文献内容与用户信息需求的匹配程度。如图 6(d)所示，在文献匹配度较高的情况下，用户浏览后下载的比例更高；而在文献匹配度较低的情况下，用户浏览后未下载的比例更高。用户的下载行为与信息需求满足程度息息相关，因此，“文献匹配度”可以作为分类的重要特征之一。实验中还对文献匹配度计算中元数据相关检索字段的附加值 β 的最佳取值进行了探索，采用穷举法令该值在[0,1]之间变化，步长为 0.1，发现 β 值对模型性能影响甚微，原因可能是数据集中存在元数据相关检索字段的记录数较少，导致基于内容相关检索字段计算的文献匹配度起主导作用。

低贡献度特征中包括“来源期刊”和“引用量”两个元数据特征，在一定程度上表明用户在文献下载决策中较少关注来源期刊的权威性和文献引用量的大小。相关研究也指出，在单篇文献层次上，下载量和引用量之间无显著相关关系[40]。“检索字段”“是否采用精确匹配”“是否为题名”的贡献度较低，在上述特征的不同取值区间内，正负样本的占比无显著差异（见图 6(g)、图 6(h)、图 6(j)）。原因可能是上述查询式构造特征在区分查询意图类型上存在不足，因此未来还需进一步探索查询意图的精确识别。

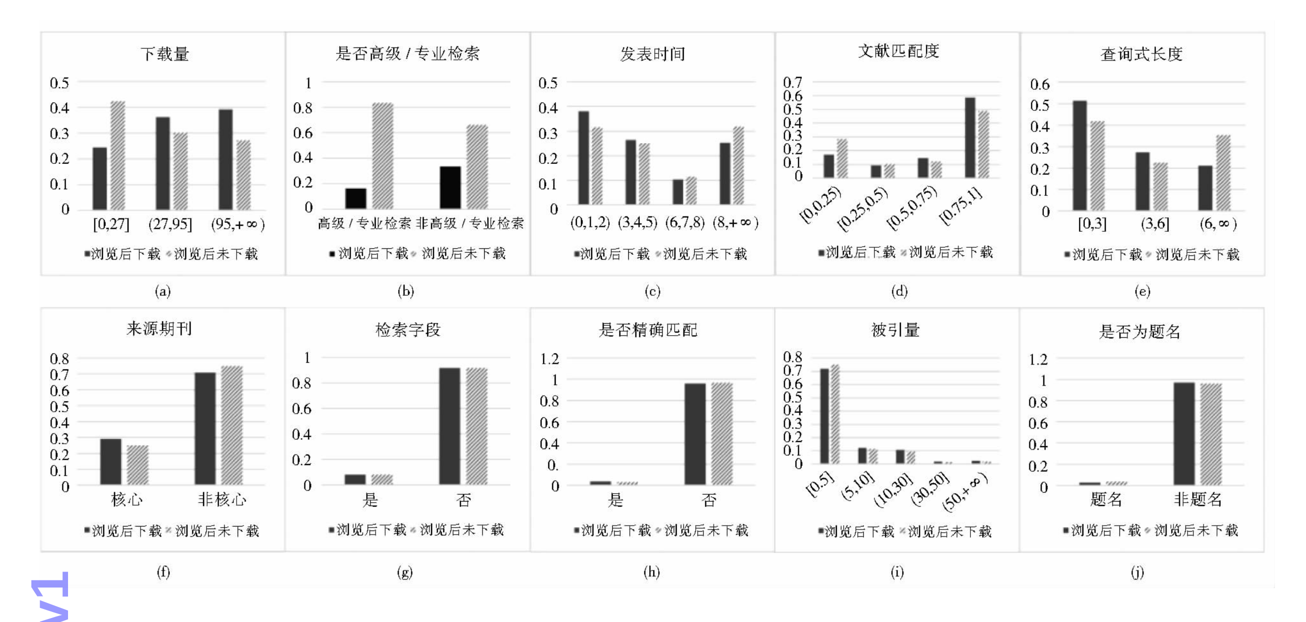


图6　查询相关性特征不同取值区间内正负样本占比分布

4　结论

笔者在建立用户学术文献下载行为的多维特征体系的基础上,采用机器学习算法分别构建了基于查询相关性和基于用户行为的学术文献下载预测子模型,并对两类模型进行加权融合。在万方用户行为日志数据集上的实验结果表明,随机森林算法的性能最佳,混合分类器能提升分类效果;基于用户行为的子分类相较于基于查询相关性的子分类器占有更高权重;“下载量”“是否采用专业/高级检索”和“发表时间”是影响用户下载行为的重要因素。笔者提出的融合多维特征的学术文献下载行为预测模型取得了良好的效果,实际应用时可作为学术检索系统重排序模块的组成部分:在检索系统根据用户查询式返回相关文献集合后,该模型基于用户行为数据和文献数据提取相应特征,通过训练好的分类器预测用户会进行下载的文献列表,最后将该列表中的文献排在检索结果页面的靠前位置,从而提高科研用户学术检索效率。

本文存在以下不足:①仅用查询式特征间接反映查询意图,未构建查询意图识别模型来明晰科研用户的检索需求,未来可考虑构建查询意图识别模型并与下载预测模型相结合;②由于数据集的局限性,无法区分机构用户中的个体科研用户,因此未引入会话分析。在未来研究中将引入会话序列,分析用户在学术检索过程中内容偏好和行为偏好的时序演化,从而对学术文献下载行为预测模型进行改进和创新。

参考文献:

[1] 熊泽泉, 段宇锋. 论文早期下载量可否预测后期被引量?——以图书情报领域期刊为例[J]. 图书情报知识, 2018(4): 32－42.

[2] 谢娟, 龚凯乐, 成颖, 等. 论文下载量与被引量相关关系的元分析[J]. 情报学报, 2017, 36(12): 1255－1269.

[3] 王超. 期刊论文被引量与下载量关系研究[J]. 情报探索, 2020(6): 33－39.

[4] 张维, 代国强. 国内高校图书馆数据库过量下载的特点及对策分析[J]. 办公自动化, 2016, 21(14): 25－27.

[5] 张敏, 张磊. 数字图书馆电子资源过量下载意愿的使能因素和抑能因素平衡研究[J]. 图书馆学研究, 2016(16): 51－57, 69.

[6] 张敏, 张磊. 数字图书馆电子资源过度下载意愿的影响因素研究——基于任务驱动与惩罚抑制的双重情境[J]. 图书情报工作, 2016, 60(7): 116－122.

[7] 孙利. 广州大学城大学生下载数据库论文行为研究[J]. 图书情报导刊, 2016, 1(11): 150－153.

[8] 刘颖. 基于 ARIMA 模型和神经网络对论文下载量进行预测[D]. 大连: 大连理工大学, 2015.

[9] LI X, DE RIJKE M. Characterizing and predicting downloads in academic search[J]. Information processing & management, 2019, 56(3): 394－407.

[10] 张海涛, 张枭慧, 魏萍, 等. 网络用户信息检索行为研究进展[J]. 情报科学, 2020, 38(5): 169－176.

[11] BRODER A. A taxonomy of Web search[J]. SIGIR forum, 2002, 36(2): 3－10.

[12] DOU Z, SONG R, WEN J. A large-scale evaluation and analysis of personalized search strategies[C]//Proceedings of the 16th inter-

national conference on World Wide Web. New York: ACM, 2007: 581 –590.

[13] KHABSA M, WU Z, GILES C L. Towards better understanding of academic search[C]//Proceedings of the 16rh ACM/IEEE-CS on joint conference on digital libraries. New York: ACM, 2016: 111 –114.

[14] 张晓娟. 信息类、导航类与事务类查询个性化潜力的对比析究[J]. 数字图书馆论坛, 2017(9): 35 –41.

[15] 吴丹, 孙浩东. 移动图书馆 WAP 和 APP 用户检索行为比较分析[J]. 图书情报工作, 2016, 60(18): 14 –20.

[16] LI X, SCHIJVENAARS B J A, DE RIJKE M. Investigating queries and search failures in academic search[J]. Information processing & management, 2017, 53(3): 666 –683.

[17] WILSON T D. Human information behavior[J]. Informing science, 2000, 3(2): 49 –56.

[18] 王建冬, 王继民. 基于日志挖掘的高校用户期刊数据库检索行为研究[J]. 北京大学学报(自然科学版), 2012, 48(1): 29 –36.

[19] 楼海萍, 潘杏梅, 方红, 等. 我国学术论文下载指标研究综述[J]. 图书馆研究与工作, 2018(10): 50 –55.

[20] 郭强, 赵瑾, 刘思源, 等. 科技论文下载次数的统计性质研究[J]. 情报科学, 2009, 27(5): 690 –694.

[21] GARFIELD E. Fortnightly review: How can impact factors be improved? [J]. BMJ, 1996, 313(7054): 411 –413.

[22] 赵一权, 王振民, 熊文炳, 等. 科学论文的下载与引用关系研究:以 ACM 数字图书馆为例[J]. 中国科技期刊研究, 2014, 25(6): 818 –823.

[23] 赵星. 学术文献用量级数据 Usage 的测度特性研究[J]. 中国图书馆学报, 2017, 43(7): 44 –57.

[24] 杨莉, 熊泽泉, 段宇锋. 基于分位数回归的期刊论文被引量预测研究[J]. 情报科学, 2019, 37(10): 60 –66.

[25] 牛昱昕, 宗乾进, 袁勤俭. 开放存取论文下载与引用情况计量研究[J]. 中国图书馆学报, 2012, 38(4): 119 –127.

[26] O' LEARY D. On the relationship between citations and appearances on "top 25" download lists in the international journal of accounting information systems[J]. International journal of accounting information systems, 2008, 9(1): 61 –75.

[27] 徐文贤, 陈雪梅. 高校图书馆数据库过量下载行为研究[J]. 图书馆理论与实践, 2014(11): 20 –23.

[28] BARKAN O, KOENIGSTEIN N. Item2vec: neural item embedding for collaborative filtering[C]//2016 IEEE 26th international workshop on machine learning for signal processing. Piscataway:IEEE, 2016: 1 –6.

[29] 杨书新, 徐慧琴, 谭伟. 结合查询相关性的关键词查询排序方法[J]. 计算机工程与设计, 2013, 34(9): 3136 –3140.

[30] 吴丽华, 罗云锋, 张宏斌. 信息检索模型及相关性算法的研究[J]. 情报杂志, 2006(12): 25 –27.

[31] 张李义, 张然. 技术接受模型(TAM)关键变量前因分析[J]. 信息资源管理学报, 2015, 5(2): 11 –20.

[32] 王海涛, 谭宗颖, 陈挺. 论文被引频次影响因素研究——兼论被引频次评估科研质量的合理性[J]. 科学学研究, 2016, 34(2): 171 –177.

[33] 沈敏, 杨新涯, 王楷. 基于机器学习的高校图书馆用户偏好检索系统研究[J]. 图书情报工作, 2015, 59(11): 143 –148.

[34] 陆伟, 周红霞, 张晓娟. 查询意图研究综述[J]. 中国图书馆学报, 2013, 39(1): 100 –111.

[35] BELKIN N J, KELLY D, KIM G, et al. Query length in interactive information retrieval[C]//Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2003: 205 –212.

[36] TANG J, WANG K. Personalized Top-N sequential recommendation via convolutional sequence embedding[C]// The eleventh ACM international conference. New York: ACM, 2018.

[37] WANG J, HUANG P, ZHAO H, et al. Billion-scale commodity embedding for e-commerce recommendation in alibaba[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. New York: ACM, 2018: 839 –848.

[38] ZHANG W, DU Y, YOSHIDA T, et al. DeepRec: a deep neural network approach to recommendation with item embedding and weighted loss function[J]. Information sciences, 2019, 470 (2019): 121 –140.

[39] "慧源共享"全国高校开放数据创新研究大赛组委会. "慧源共享"全国高校开放数据创新研究大赛 – 参赛作品提交须知[EB/OL]. [2020 –07 –01]. http://hdl. handle. net/20. 500. 12291/10232 V2 [Version].

[40] 陆伟, 钱坤, 唐祥彬. 文献下载频次与被引频次的相关性研究——以图书情报领域为例[J]. 情报科学, 2016, 34(1): 3 –8.

作者贡献说明:

谢豪:模型构建,实验代码编写,论文撰写与修订;
吴雪华:特征体系构建,模型构建,论文撰写与修订;
陈茜:特征体系构建,模型构建,论文撰写与修订;
唐晶:模型构建,论文撰写与修订;
白云:模型构建,论文撰写与修订;
毛进:提出研究思路和论文修改建议。

谢豪，吴雪华，陈茜，等.融合多维特征的学术文献下载行为预测研究[J].图书情报工作,2021,65(12):112－121.

Predicting Download Behavior of Academic Literature Based on Multi-dimensional Features

Xie Hao　Wu Xuehua　Chen Xi　Tang Jing　Bai Yun　Mao Jin
Center for Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] The behavior of academic literature downloading is an essential step in the process of academic retrieval. Predicting download behavior of academic literature is conducive to the in-depth understanding of the retrieval behavior of researchers, and provides a basis for optimizing retrieval results of academic resource retrieval platforms and restructuring ranking, to improve the retrieval function and service quality of retrieval system. [**Method/process**] This paper constructed a multi-dimensional feature system of researchers' academic literature download behavior, and proposed two sub-classifiers based on query relevance and user behavior respectively relying on machine learning algorithms. A weighted strategy was adopted to construct a hybrid model of download behavior prediction of academic literature. [**Result/conclusion**] The experiment results show that the Random Forest algorithm achieves the best performance in both classifiers. Compared to the model trained with only query relevance features, the accuracy of the hybrid model is increased by 2.3%, and the F1 value is increased by 1.3%. The sub-classifiers based on user behavior have higher weights in the hybrid model. "downloads""whether professional/advanced search is used"and "published time" make a significant contribution to the academic literature download prediction task.

Keywords: literature download prediction　multi-dimensional features　machine learning　hybrid model